Dates: 22.09.24 & 23.09.24

Machine Learning

Assignment – 5

Answers

1. R-squared is generally a better measure of the goodness of fit for a regression model than the residual sum of squares (RSS).

R-squared, denoted as R^2, is a statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. It is dimensionless and ranges from 0 to 1, where a value closer to 1 indicates a better fit. R^2 is calculated using the formula:

R^2 = 1 - RSS/TSS

where RSS is the residual sum of squares, and TSS is the total sum of squares.

The reason why R^2 is often preferred over RSS as a measure of goodness of fit is due to its standardized nature:

1. Scalability: R^2 is scale-invariant whereas RSS is affected by the scale of the dependent variable. This makes R^2 a better choice when comparing models fitted on different scales.

2. Interpretability: R^2 has an intuitive interpretation as the proportion of variance explained, which is easier to understand than the sum of squared residuals.

3. Benchmarking: R^2 provides a clear benchmark.

4. Adjustment for model complexity: Adjusted R^2 takes into account the number of predictors in the model, which helps in assessing whether the addition of a new predictor really improves the model or is just adding complexity without significantly improving the fit.

2. TSS is the total sum of squares. TSS represents the total variance in the dependent variable.

The sum of squares total (SST) or the total sum of squares (TSS) is the sum of squared differences between the observed dependent variables and the overall mean. Think of it as the dispersion of the observed variables around the mean—similar to the variance in descriptive statistics. It is calculated as follows:

TSS = $\sum_{i=1}^{n}$ (y_i – y bar)^2

where

y_i - observed dependent variable

y bar - mean of the dependent variable

The sum of squares due to regression (SSR) or explained sum of squares (ESS) is the sum of the differences between the predicted value and the mean of the dependent variable. In other words, it describes how well our line fits the data. It is calculated as follows:

ESS = $\sum_{i=1}^{n}$ ($\hat{y}_i$ - y bar)^2

where

$\hat{y}_i$ - the predicted value of the dependent variable

y bar - mean of the dependent variable

RSS is the residual sum of squares or the sum of squares error (SSE). It is the sum of the squared differences between the observed actual outcomes and the outcomes predicted by the regression model. It is calculated as:

RSS = $\sum_{i=1}^{n}$ $(y_i - \hat{y}_i)^2$

where y_i is the actual value and $\hat{y}_i$ is the predicted value from the model for the i-th observation.

Mathematically, SST = SSR + SSE or TSS = ESS + RSS.

3. Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting and underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

4. The Gini Index is also known as Gini impurity. It is a measure of how mixed or impure a dataset is. The Gini impurity ranges between 0 and 1, where 0 represents a pure dataset and 1 represents a completely impure dataset. It is represented by the following formula –

I_G = 1 - $\sum_{j=1}^{c} p^2 j$

where p_j is proportion of the samples that belongs to class c for a particular node.

The decision tree algorithm CART (Classification and Regression Tree) uses the Gini method to create split points.

5. Yes. Decision trees are a popular and powerful method for data mining, as they can handle both numerical and categorical data, and can easily interpret the results. However, decision trees can also suffer from overfitting, which means that they learn too much from the training data and fail to generalize well to new data. Overfitting can lead to poor performance, inaccurate predictions, and reduced reliability. There are several techniques that can help prevent overfitting in decision trees, such as pruning, regularization, and ensemble methods. Regularization is the process of adding some constraints or penalties to the tree growth, such as limiting the depth, the number of nodes, or the

minimum samples required for a split. Regularization can help reduce the complexity and variance of the tree, and avoid overfitting.

6. Ensemble technique or learning is a machine learning paradigm where multiple models (often called “weak learners”) are trained to solve the same problem and combined to get better results. The main hypothesis is that when weak models are correctly combined we can obtain more accurate and/or robust models.

7. Difference between Bagging and Boosting techniques – Both Bagging and Boosting are meta-algorithms that aim at combining weak learners.

Bagging, that often considers homogeneous weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process.

Boosting, that often considers homogeneous weak learners, learns them sequentially in a very adaptive way (a base model depends on the previous ones) and combines them following a deterministic strategy.

8. The out-of-bag error is an error estimation technique often used to evaluate the accuracy of a random forest and to select appropriate values for tuning parameters, such as the number of candidate predictors that are randomly drawn for a split, referred to as mtry.

9. K-fold cross validation randomly splits the training data into K subsets called folds. It is easy to follow and implement. This method trains the model on a large portion of the dataset, has a good ratio of testing data points and iterates on the training and testing process multiple times. The average of the k recorded errors is called the cross validation error and will serve as the performance metric of the model.

10. Hyper parameter tuning in machine learning refers to tuning of Kernel, Regularization and Gamma. Tuning of Kernel can lead to more accurate classifiers. Regularization controls the trade-off between decision boundary and misclassification term. The value of Gamma is related to fitting of the training dataset.

11. In Gradient Descent a large learning rate might not allow it to converge to the global minimum. When the learning rate is too large, gradient descent can suffer from divergence. This means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values. In order for Gradient Descent to work, we must set the learning rate to an appropriate value. This parameter determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution.

12. Logistic regression assumes a linear relationship between the input features and the target variable. This means it may not perform that well when the relationship between the feature and outcomes is non-linear (This is where the neural network comes in).

13. Difference between Adaboost and Gradient Boosting –

In Adaptive boosting (often called "Adaboost") we try to define our ensemble models as a weighted sum of L weak learners.

$S_L(.) = \sum_{l=1}^{L} c_l * w_{l(.)}$

where c_l's are coefficients and w_l's are weak learners

Adaboost updates weights of the observations at each iteration. Weights of well classified observations decrease rapidly to weights of misclassified observations. Models that perform better have higher weights in the final ensemble model.

In Gradient boosting, the ensemble model that we try to build is also a weighted sum of weak learners.

$S_L(.) = \sum_{l=1}^{L} c_l * w_{l(.)}$

where c_l's are coefficients and w_l's are weak learners

The main difference with adaptive boosting is in the definition of sequential optimization process. The gradient boosting can be considered as a generalization of adaboost to arbitrary differentiable loss functions.

Gradient boosting updates values of the observations at each iteration. Weak learners are trained to fit the pseudo-residuals that indicate in which direction to correct the current ensemble model predictions to lower the error.

14. Bias-Variance Trade off - Bias occurs when an algorithm has limited flexibility to learn from data. Variance defines the algorithm's sensitivity to specific sets of data. An optimal model is one in which the model is sensitive to the pattern in our model, but at the same time can generalize to new data. This happens when Bias and Variance are both optimal. We call this Bias-Variance Trade off and we can achieve it in over or under fitted models by using Regression.

15. The following are three of the common kernels used with SVM:

Linear or Linear splines kernel in one dimension – it is useful when dealing with large sparse data vectors. It is often used in text categorization. The splines kernel also performs well in regression problems.

RBF or Gaussian radial basis function kernel – it is a general purpose kernel and is used when there is no prior knowledge about the data.

Polynomial kernel – it is popular in image processing. The equation is as follows:

$k(x_i, x_j) = (x_i . x_j + 1)^d$ where d is the degree of the polynomial.